US CMS Tier2 Center Prototype
In California

Introduction

Caltech and UCSD jointly propose to implement a prototype Tier2 center in September 2000. As discussed at the US CMS Collaboration Meeting in May 2000, at the recent DOE Review of Software and Computing, and during the Hoffmann Review of LHC Computing, Tier2 centers are an important part of the distributed data analysis model (comprising approximately one-half of US CMS' proposed computing capability) that has been adopted by all four LHC Collaborations. The use and management of the regional centers requires the use of "Data Grid" tools, some of which are already under development in CMS[1]. The prototype system, and the coordination of its operation with CERN and Fermilab for production and analysis, will provide a testing ground for these tools.

A cost effective candidate configuration for the prototype has already been developed, consisting of Linux rackmounted computational servers, medium scale data servers and network interfaces that have been demonstrated to provide high I/O (100 Mbyte/sec range) throughput, and a few-Terabyte RAID array as a nearline data store for simulated, reconstructed and analyzed events.

The prototype center will be located at the Caltech Center for Advanced Computing Research (CACR) and the San Diego Supercomputer Center (SDSC). This will allow us to leverage the existing large scale HPSS tape storage systems; system software installed and being developed in association with the Particle Physics Data Grid (PPDG), ALDAP and GriPhyN projects; as well as the manpower located at each of these facilities. The CALREN OC-12 (622 Mbps) California regional network and the NTON dark fiber link interconnecting CACR and SDSC will be used to test distributed center operations at speeds characteristic of future networks (0.6 – 2.5 Gbps). In the future we could also consider the use of CALREN or its successor (and/or NTON) to include more California sites, such as UC Davis, UC Riverside and UCLA.

At the January 2000 DOE/NSF review of CMS and ATLAS Software and Computing, the agencies requested that we begin to better define the roles and responsibilities of Tier2 centers relative to the Tier0 center at CERN, and the Tier1 center at Fermilab. The prototype will serve as a testbed for defining the Tier2's roles and its modes of operation. The Abilene and ESNet networks as well as the US-CERN link will be used to support these tests.

[1] Work now underway, done in collaboration with Ian Foster and Carl Kesselman as part of the PPDG, GriPhyN, and EU DataGrid Projects.

Work Plan

The work plan has three major branches:

- R&D on the distributed computing model, as summarized above. Strategies for production processing, and for data analysis, will be investigated with the help of the MONARC simulation tools
- Help with upcoming production milestones in association with the Physics Reconstruction and Selection (PRS) studies
- Startup of a prototypical US-based data analysis among the California universities in CMS.

System Configuration

The prototype system consists of two symmetric pieces, as shown in Figure 1. One located at the Caltech Center for Advanced Computing Research and the other at the San Diego Supercomputing center connected over a high speed internet link, NTON dark fiber and CALREN OC-12.

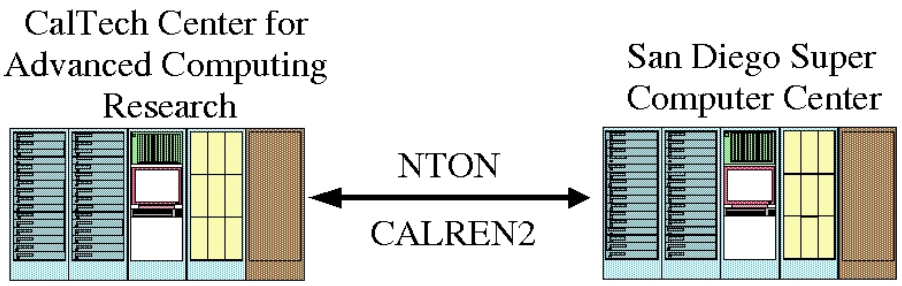

Figure 1: A schematic of the prototype Tier2 center.

Each side contains approximately 40 rack-mounted, dual CPU Pentium III LINUX nodes, a multiple terabyte RAID disk array connected to a data server, a network switch, a connection to the existing high performance storage system (HPSS), and a control monitor. Each node contains 2 hard disks, one 10 gigabyte disk for the operating system and swap space and one 30 gigabyte disk for local data cache; 512 Mbytes of 133 MHz SDRAM; 2 Pentium III 800 MHz processors; and one 100 base-T Ethernet card. The nodes are connected via a high performance network switch with approximately 50 gigabits/sec bandwidth on the backplane. This allows a sufficient number of 100 base-T Ethernet ports to interconnect the nodes and several gigabit/sec Ethernet ports to connect the switch to the RAID array, the HPSS, and the external network. Figure 2 shows the network connections of the Tier2 prototype system.

Each of the nodes is connected to the control terminal through a monitor/keyboard switch. This allows local console access to each node. The system can be administrated either locally or remotely. We plan to support a single system manager and one computer

technician as part of the Tier2 prototype project. The budget contains some initial support for this staff. Further funding would be required during the period of operation.

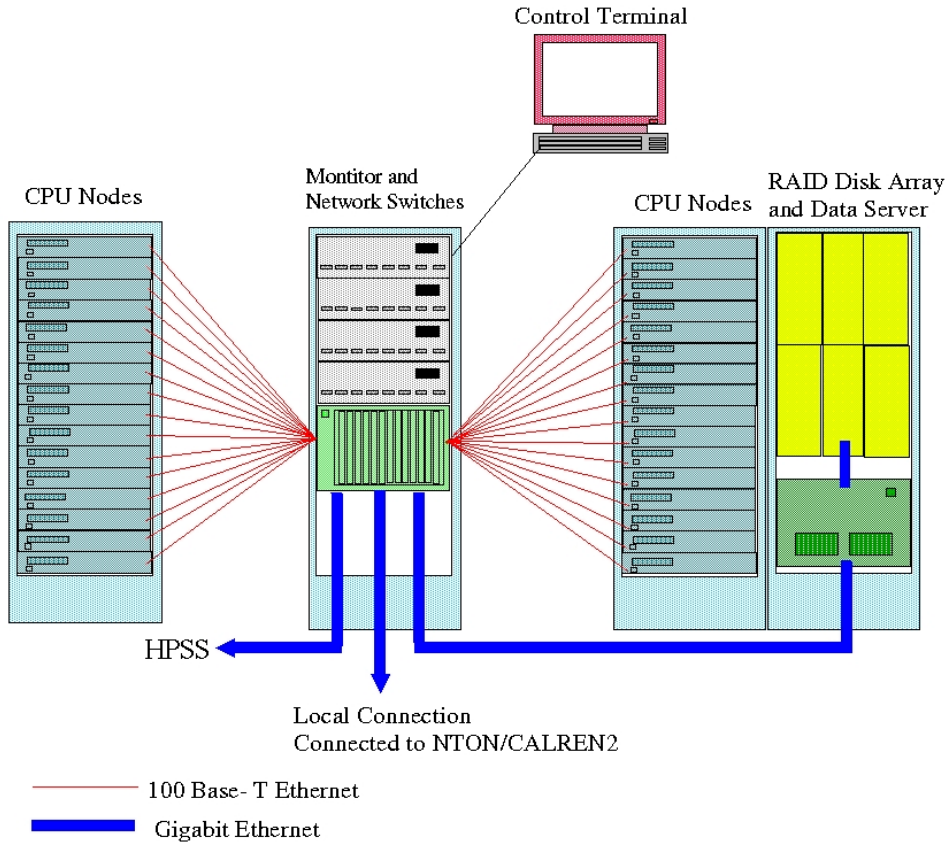

Figure 2: Network connections of the Tier2 prototype system.

Equipment to purchase for each site:

- 40 Dual Pentium III computational nodes
- A multi-terabyte RAID disk array
- 1 network switch with 100 base-T Ethernet ports and gigabit Ethernet ports with approximately 50 gigabits of bandwidth on the backplane
- 5 8-port monitor/keyboard switches
- 1 control terminal
- 3-4 equipment racks
- Network and monitor cabling

Services available from the sites:

- Multi-terabyte high performance storage system (HPSS)
- Conditioned power and cooling
- High bandwidth connections between the sites

- RAID array, managed by existing Dell 4400 data server and DLT tape library (for sending DLT copies if needed) at Caltech
- Basic collaborative infrastructure

Schedule and Deployment Plan

The schedule for selecting the vendor, system delivery and setup is as follows, assuming a September 1 go-ahead:

- Vendor selection for all system components — September 15
- Systems delivery — October 9
- Start of Operations — October 23

Leveraged Resources

As stated above, the two sites allow us to use existing hardware, local infrastructure, and high speed network connections. The HPSS available at each site is of particular interest. There are additional telecom upgrades and initiatives planned in California and in the UC system that will improve our ability to work together in the future.

In addition, Caltech, UCSD, CACR and SDSC personnel are available an interested in developing the prototype. This includes:

Name	Affiliation	Tasks	%
Julian Bunn	Caltech	ODBMS	20%
Harvey Newman	Caltech	Overall Direction	10%
Koen Holtman	Caltech	PPDG and scheduling	50%
Iosif Legrand	US CMS S&C	simulations and network monitoring tools	80%
Vladimir Litvin	US CMS S&C	ORCA and CMS environment support	25%
Takako Hickey	Caltech	ALDAP and scheduling methods	50%
Mehnaz Hafeez	Caltech	Globus infrastructure	40%
Asad Samar	Caltech	Globus infrastructure	30%
Philippe Galvez	Caltech	network measurements	30%
James Patton	CACR	Network setup and NTON measurements	
Jan Lindheim	CACR	data server configuration	
Ian Fisk	UCSD	ORCA, Physics tools	25%
Jim Branson	UCSD	Physics Tools	10%
Reagan Moore	SDSC	grid data	
Mike Vildibill	SDSC	grid data handling system	
Phil Andrews	SDSC	grid data handling system	

Role of Other Universities in California and Elsewhere

All of the CMS member universities in California will be connected at high speed by the Calren-2 project. We expect that UC Davis, UCLA, and UC Riverside will be prototype

users of this prototype center, in addition to CMS members at UCSD and Caltech. All of these universities will have significant software and physics efforts next year, when the center is operating. We will also provide support for users from other US regions, if that turns out to be appropriate in the presence of other resources at FNAL and elsewhere.

It may be possible, later, to enlarge the center by including resources at the other California universities. UC Riverside is interested in this possibility.

Budget (revised 10/3/00)

We have consulted several vendors to get price estimates. As of Fall 2000, 800 MHz dual processor nodes with 512 MB of 133 MHz memory and two disks in a good 2U rack mountable enclosure are available for approximately $ 2500. At this time, we assume a standard RAID disk array and good Gigabit/Fast Ethernet switches.

Tier2 Center Prototype: Revised Budget (10/3/00)

This budget takes the UC Davis cost sharing for local Tier3 equipment (at no cost to the project) explicitly into account.

• 80 Dual CPU + Disk Linux Nodes	$ 200 k
• Sun Data Server with RAID Array	$ 30 k [*]
• Tape Library	$ 20 k
• LAN Switches	$ 50 k
• Collaborative Infrastructure Upgrades	$ 10 k
• Installation and Infrastructure	$ 30 k
• Net Connect to Abilene	$ 0 k
• Tape Media and Consumables	$ 10 k
• Staff (Ops and System Support)	$ 50 k
• Total Tier2 Estimated Cost (First Year)	$ 400 k
• Tier3 Client and Tier2-host systems at UC Davis: CPUs, Disks, Video System Upgrade, Net interfaces	$ 30 k
• UCSD cost sharing	$ -50 k
• UC Davis cost sharing	$ -30 k
• Net Cost	$ 350 k

Proposed Budget Allocations

• UCSD	$ 175 k
• Caltech	$ 115 k
• UC Davis (for equipment to be sited at Caltech)	$ 60 k

[*] Funding for this RAID array is partly existing from other sources at Caltech.